STAT1378 Presentation

Ze Hong Zhou

03/11/2021

Research Questions

1) Do hentai animes have a lower average score compared to non-hentai animes?
2) Is the average score of animes the same across all release dates?

Overview

- Data set
- Question 1 & Question 2
  - Definitions
  - Hypotheses and Test Statistic
  - Assumptions
  - Conducting the Test
  - Discussion and Conclusion
- References

Data set

- The data set was scraped from MyAnimeList (MAL) (Valdivieso (2020))
- MAL is currently the largest anime database

Question 1: Do hentai animes have a lower average score compared to non-hentai animes?

Definitions

- The Western definition of hentai is a genre of Japanese anime and manga that contains pornography.
- We will use H to denote the hentai anime population and N to denote the non-hentai anime population.

Definitions (continued)

- In MAL, "Hentai" is a tag under "Genres." This is the what we will be using to classify if an anime is hentai or not.

Figure 1: The information section of a hentai anime in MAL.

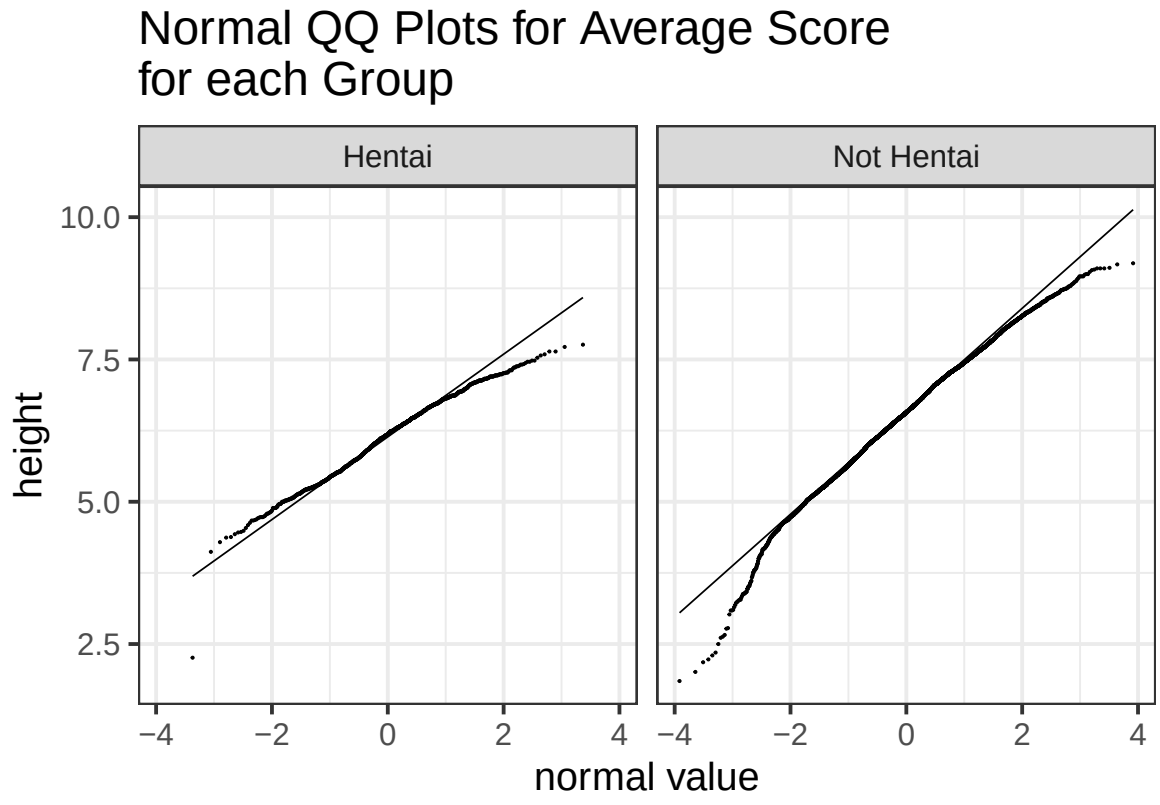
Hypotheses and Test Statistic

- We are testing:
  - H_0: $\mu_H - \mu_N = 0$ against H_1: $\mu_H - \mu_N < 0$.
- Test statistic:
  - $\tau = \frac{\overline{x}_H - \overline{x}_N}{\sqrt{\frac{s_H^2}{n_H} + \frac{s_N^2}{n_N}}} = \frac{\overline{x}_H - \overline{x}_N}{\sqrt{\frac{s_H^2}{1330} + \frac{s_N^2}{11091}}}$
- If H_0 is true AND assumptions are satisfied:
  - $\tau \dot{\sim} t_{1992} \doteq Z$

Assumptions

- Data from MAL is a random sample from each group population.
  - We exclude observations with an unknown score or genre.
- Observations are therefore independent of each other within and across each group.

Assumptions (continued)



- The average score variable for each group seems to be normally distributed since the QQ plots mostly follow a straight line.

Conducting the Test

- Since the statistical test assumptions are satisfied, we can now go ahead with the test.
- We use the function `t.test` in the `stats` package in R (R Core Team (2021)).

Conducting the Test (continued)

Table 1: summary of the Welch two sample test

τ_{obs}	-21.1
95% confidence interval for $\mu_H - \mu_N$	$[-\infty$, -0.386]
degrees of freedom	1992
p-value	8.8e-90

As seen in table 1, the p-value $<<$ 5%, so we reject H_0.

Discussion and Conclusion

- Our aim was to determine if the average score of hentai animes is lower than that of non-hentai animes.
- After testing the assumptions for the Welch two sample t-test and conducting it, we conclude that it is indeed lower.

Question 2: Is the average score of animes the same across all release dates?

Definitions

- The release date of an anime is the date that the first episode aired.
- MAL contains information on the air dates of animes under "Aired."

Definitions (continued)

In figure 2, the release date of "Shinsekai yori" is Sep 29, 2012.

Information

Type: TV

Episodes: 25

Status: Finished Airing

Aired: Sep 29, 2012 to Mar 23, 2013

Premiered: Fall 2012

Broadcast: Saturdays at 00:30 (JST)

Producers: Aniplex, TV Asahi, Pony Canyon

Licensors: Sentai Filmworks

Studios: A-1 Pictures

Source: Novel

Genres: Drama, Horror, Mystery, Sci-Fi, Supernatural

Theme: Psychological

Duration: 22 min. per ep.

Rating: R - 17+ (violence & profanity)

Figure 2: The information section of Shinsekai yori in MAL.

Hypotheses and Test Statistic

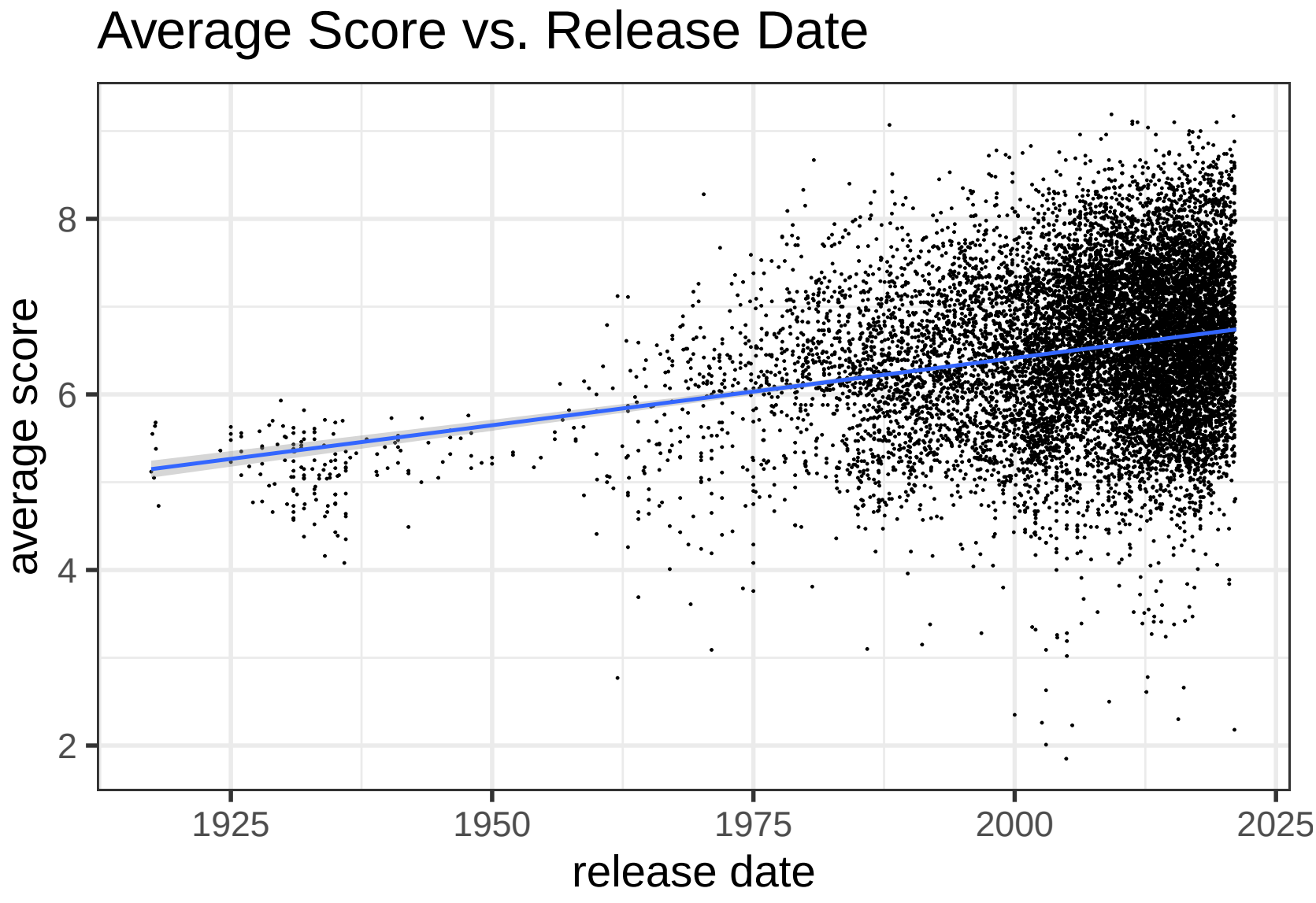
Let the independent variable, X, be release date, and the dependent variable, Y, be average score.

- We are testing:
  - H_0: $\beta = 0$ against H_1: $\beta \neq 0$.
- Test statistic:
  - $\tau = \frac{\hat{\beta}}{s_{Y|X}/\sqrt{S_{XX}}}$
- If H_0 is true AND assumptions are satisfied:
  - $\tau \sim t_{12412} \doteq Z$

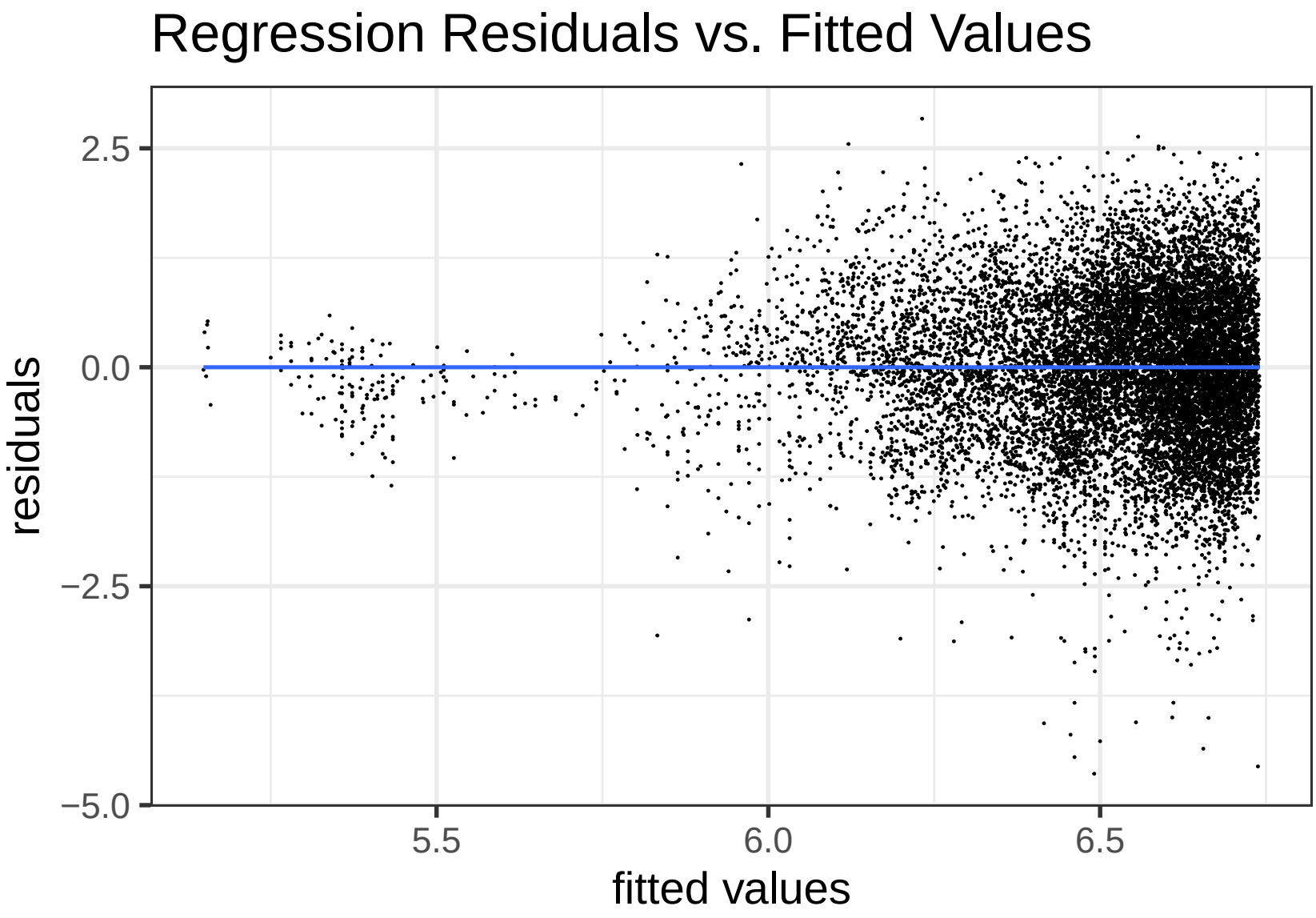
Assumptions

- Data from MAL is a random sample from the anime population.
  - We exclude observations with an unknown score or release date.
- Observations are therefore independent of each other.

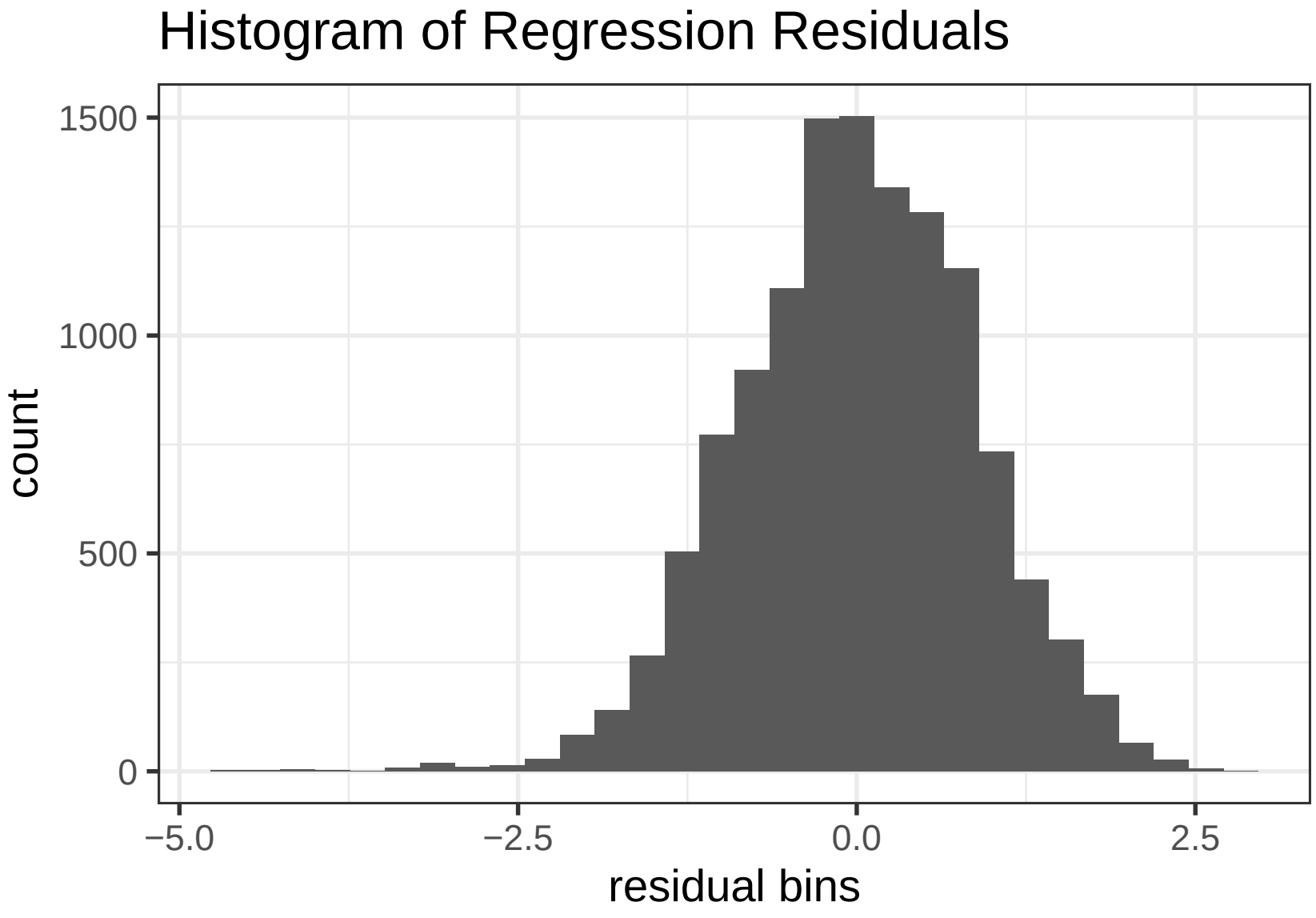
Assumptions (continued)



Assumptions (continued)



Assumptions (continued)



Assumptions (continued)

- There seems to be a linear trend between average score and release date.
- For any value of the fitted value, the residuals seem to be normally distributed with a constant variance.

Conducting the Test

- Since the statistical test assumptions are satisfied, we can now go ahead with the test.
- We use the function `lm` in the `stats` package in R (R Core Team (2021)).

Conducting the Test (continued)

Table 2: summary of the linear regression coefficient t-test

$\hat{\beta}$	0.015
95% confidence interval for β	[0.014, 0.016]
t-value	28
degrees of freedom	12412
p-value	3.1e-167

As seen in table 2, the p-value $<<$ 5%, so we reject H_0.

Discussion and Conclusion

- Our aim was to determine if the average score of animes is the same across all release dates.
- After testing the assumptions for the linear regression coefficient t-test and conducting it, we conclude that it is not the same across release dates, but rather it increases by about 0.015 each year.

Thank you!

Thank you for your attention!

I hope that this has answered all your anime curiosities.

References

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Valdivieso, Hernan. 2020. "Anime Recommendation Database." Santiago, Chile. https://github.com/Hernan4444/MyAnimeList-Database.